AStA (2007) 91: 279–290
DOI 10.1007/s10182-007-0035-0

ORIGINAL PAPER

On the inefficiency of propensity score matching

Markus Frölich

Received: 21 August 2006 / Accepted: 24 April 2007 / Published online: 21 September 2007

Summary *Propensity score matching* is now widely used in empirical applications for estimating treatment effects. Propensity score matching (PSM) is preferred to matching on X because of the lower dimension of the estimation problem. In this note, however, it is shown that PSM is *inefficient* compared to matching on X. Hence, matching on X should be considered as a serious alternative.

Keywords Propensity score matching · Average treatment effect · Efficiency

1 Introduction

Propensity score matching is now widely used in empirical applications for estimating treatment effects.[1] Propensity score matching (PSM) is preferred to matching on X because of the lower dimension of the estimation problem. For PSM only one-dimensional nonparametric regression is needed, whereas matching on X requires nonparametric regression of dimension $\dim(X)$, which is more difficult to implement. In this note, it is shown, however, that PSM is *inefficient* compared to matching on X when estimating average treatment effects.[2] Hence, matching estimation on X should be considered as a serious alternative. (Throughout this paper, the expression 'matching on X' is used in the terminology of Heckman et al. (1998b) as an average of a nonparametric regression estimator. It is thus different from 'exact matching on X',

M. Frölich (✉)
SIAW, Universität St. Gallen, Bodanstr. 8, 9000 St. Gallen, Switzerland
e-mail: markus.froelich@unisg.ch

[1] See e.g., Black and Smith (2004), Frölich (2004), Gerfin and Lechner (2002), Heckman et al. (1998a), Frölich et al. (2004), Larsson (2003), Lechner (1999, 2002), Sianesi (2004) or Smith and Todd (2005).

[2] Hahn (1998) showed this for the case of experimental data, where the propensity score is *constant* and not affected by the X variables.

which could be considered as the limit case for a bandwidth value of zero, but is generally not efficient.)

Consider a binary treatment (e.g., participation in a training programme) and denote the potential outcomes as Y_i^0 and Y_i^1, where Y_i^1 is the outcome when participating in the treatment whereas Y_i^0 is the outcome when not participating. Let $D_i \in \{0, 1\}$ denote whether an individual received treatment or did not. We are interested in the average treatment effect (ATE) and the average treatment effect on the treated (ATET):

$$\begin{aligned} \text{ATE} &= E\left[Y^1 - Y^0\right], \\ \text{ATET} &= E\left[Y^1 - Y^0 \middle| D = 1\right]. \end{aligned}$$

If one observes all variables X that affected the potential outcomes as well as the treatment participation decision, the potential outcomes are independent of treatment given X:

$$Y^0, Y^1 \perp\!\!\!\perp D|X\,. \tag{1}$$

If, in addition, treatment assignment was not deterministic in that $0 < \Pr(D = 1|X) < 1$ a.s. the treatment effects are identified and can be estimated as

$$\begin{aligned} \widehat{\text{ATE}} &= \frac{1}{n}\sum_{i=1}^{n}\left(\hat{m}_1(X_i) - \hat{m}_0(X_i)\right), \\ \widehat{\text{ATET}} &= \frac{1}{n_1}\sum_{i=1}^{n}\left(Y_i - \hat{m}_0(X_i)\right)D_i\,, \end{aligned} \tag{2}$$

where $n_1 = \sum D_i$ is the number of treated and $\hat{m}_d(x)$ is a nonparametric regression estimator, e.g., local linear, of the conditional mean functions $m_d(x) = E[Y|X = x, D = d]$. These estimators are called matching estimators by Heckman et al. (1998b).[3]

Propensity score matching is a very widely used alternative to matching on X. Rosenbaum and Rubin (1983) have shown that the conditional independence assumption (1) implies

$$Y^0, Y^1 \perp\!\!\!\perp D\,|p(X), \tag{3}$$

where $p(x) = \Pr(D = 1|X = x)$ is the propensity score.[4] Hence, the treatment effects can be estimated by matching with X replaced by $p(X)$ in (2) and $m_d(x)$ replaced by $\mathfrak{m}_d(\rho) = E[Y|p(X) = \rho, D = d]$. If the propensity score is known, PSM avoids the high-dimensional nonparametric regression of $m_d(x)$. If the propensity score is unknown, it is usually estimated parametrically. Implementation of PSM is therefore very simple and convenient in practice. In a first step, the propensity score is often

[3] As mentioned before, this should not be confused with *exact* matching on X, where only treated and non-treated units with exactly the same values of X are compared. Exact matching would have a lower bias but a larger variance. Exact matching will often be impossible if X contains a (moderately) large number of covariates or if it contains continuous variables.

[4] Frölich (2007) examines PSM under weaker assumptions.

estimated by logit regression. The estimated propensity scores are then plugged into a *one-dimensional* nonparametric regression estimator to obtain the ATE or ATET. Matching on X, on the other hand, requires higher dimensional nonparametric regression, which is harder to implement and computationally more demanding.

Matching on X, however, has often the advantage of being more efficient than PSM, as discussed below.

2 Inefficiency of propensity score matching

The following proof is based on comparing the semiparametric variance bounds for matching on X and for PSM. The semiparametric variance bound is the smallest variance that could be obtained by a consistent, unbiased, normally regular estimator. It is the counterpart to the well known Cramer–Rao bound for a nonparametric context, i.e., where no information on parametric functional forms is available. The estimators introduced in the preceding section are semiparametric in the sense that all components are nonparametric but the final object of interest, the ATE or ATET, is a scalar, i.e., a one-dimensional object, which can be estimated at $\sqrt{N}$ rate under certain regularity conditions. A semiparametric variance bound often exists for many semiparametric problems if $\sqrt{N}$ consistent estimation of the final object of interest is possible. If such a *semiparametric variance bound* exists, no semiparametric estimator can have lower variance than this bound, and any estimator that attains this bound is semiparametrically efficient.[5] Suppose the propensity score is known, these bounds are:

$$\mathcal{V}^X_{\text{ATE}} = E\left[\frac{\sigma_1^2(X)}{p(X)} + \frac{\sigma_0^2(X)}{1-p(X)} + (m_1(X) - m_0(X) - \text{ATE})^2\right], \tag{4}$$

$$\mathcal{V}^X_{\text{ATET}} = \frac{1}{\Gamma^2} E\left[p(X)\sigma_1^2(X) + p^2(X)\frac{\sigma_0^2(X)}{1-p(X)} + p^2(X)\,(m_1(X) - m_0(X) - \text{ATET})^2\right]$$

where $\Gamma = \Pr(D=1)$ and $m_d(x) = E[Y|X=x, D=d]$ and $\sigma_d^2(x) = Var[Y|X=x, D=d]$. The superscript X in $\mathcal{V}^X_{\text{ATE}}$ and $\mathcal{V}^X_{\text{ATET}}$ indicates that matching is on X.

The bounds when matching on the propensity score, denoted by the random variable P, are

$$\mathcal{V}^{\text{PSM}}_{\text{ATE}} = E\left[\frac{\acute{\sigma}_1^2(P)}{P} + \frac{\acute{\sigma}_0^2(P)}{1-P} + (\mathfrak{m}_1(P) - \mathfrak{m}_0(P) - \text{ATE})^2\right], \tag{5}$$

$$\mathcal{V}^{\text{PSM}}_{\text{ATET}} = \frac{1}{\Gamma^2} E\left[P\acute{\sigma}_1^2(P) + P^2\frac{\acute{\sigma}_0^2(P)}{1-P} + P^2 \cdot (\mathfrak{m}_1(P) - \mathfrak{m}_0(P) - \text{ATET})^2\right]$$

[5] Semiparametric efficiency bounds were introduced by Stein (1956) and developed by Koshevnik and Levit (1976), Pfanzagl and Wefelmeyer (1982), Begun et al. (1983) and Bickel et al. (1993). See also the survey of Newey (1990) or Newey (1994). Hahn (1998) derived the semiparametric variance bounds when matching on X.

where $\mathfrak{m}_d(\rho) = E[Y|P = \rho, D = d]$ and $\hat{\sigma}_d(\rho) = Var[Y|P = \rho, D = d]$. The superscript PSM in $\mathcal{V}^{\text{PSM}}_{\text{ATE}}$ and $\mathcal{V}^{\text{PSM}}_{\text{ATET}}$ indicates that matching is on the propensity score. The following theorem shows that PSM is inefficient vis-a-vis matching on X.

Theorem 1 (Inefficiency of propensity score matching). *The difference between the asymptotic variances of PSM and matching on X is non-negative*

$$\mathcal{V}^{\text{PSM}}_{\text{ATE}} - \mathcal{V}^{X}_{\text{ATE}} \geq 0\,,$$
$$\mathcal{V}^{\text{PSM}}_{\text{ATET}} - \mathcal{V}^{X}_{\text{ATET}} \geq 0\,.$$

Generally, the difference is strictly positive unless the support of P contains only values where either both variances $V_1(\rho)$ and $V_0(\rho)$ are zero, with $V_d(\rho) \equiv Var(m_d(X)|p(X) = \rho)$, or where $\sqrt{\frac{V_1(\rho)}{V_0(\rho)}} = \frac{\rho}{1-\rho}$ and $corr(m_1(X), m_0(X)|p(X) = \rho) = -1$. In this rather special case, $\mathcal{V}^{\text{PSM}}_{\text{ATE}} = \mathcal{V}^{X}_{\text{ATE}}$ and $\mathcal{V}^{\text{PSM}}_{\text{ATET}} = \mathcal{V}^{X}_{\text{ATET}}$.

3 Implementation issues

The implementation of the matching estimator requires nonparametric regression estimators $\hat{m}_d(x)$ of the conditional mean functions $m_d(x) = E[Y|X = x, D = d]$ that are plugged into:

$$\widehat{\text{ATE}} = \frac{1}{n}\sum_{i=1}^{n}\left(\hat{m}_1(X_i) - \hat{m}_0(X_i)\right). \tag{6}$$

Since X often has to contain many variables to make the conditional independence assumption (1) plausible, *higher dimensional* nonparametric regression is required, which may not only be more difficult to implement but can be computationally demanding. It is worthwhile to point out that although the 'curse of dimensionality' will lead to low precision in the estimated $\hat{m}_d(x)$, the estimators of ATE and ATET, which are averages of $\hat{m}_d(x)$, can achieve $\sqrt{n}$ rate irrespective of the dimension of X. In this sense, the curse of dimensionality does not apply to estimators of ATE and ATET. (Nevertheless, the dimension of X does still matter in that stronger regularity conditions are required and that computation time increases with $\dim(X)$.)

A large number of potential nonparametric regression estimators for $\hat{m}_d(x)$ are available. The following illustrations focus on $\hat{m}_0(x)$ where only the observations with $D = 0$ are used. The estimation of $\hat{m}_1(x)$ is analogous, using the $D = 1$ observations instead. Hence, $\hat{m}_1(x)$ and $\hat{m}_0(x)$ are estimated from distinct subsamples and are then combined in (6).

A kernel regression estimator of $\hat{m}_0(x)$ is

$$\hat{m}_0(x) = \frac{\sum_{j:D_j=0} Y_j \cdot \mathbf{K}_H\left(X_j - x\right)}{\sum_{j:D_j=0} \mathbf{K}_H\left(X_j - x\right)}\,,$$

where only the observations with $D = 0$ are used. $\mathbf{K}_H(\cdot)$ is a multivariate kernel function that assigns weights to the observations j according to their distance from x. The size of the local neighbourhood is determined by a matrix of bandwidths H. Observations outside this local neighbourhood receive a kernel weight of zero.

Often a fixed kernel function with the same bandwidths H for every value of x is used. Alternatively the bandwidths may be adjusted e.g., to the density $f_{X|D=0}(x)$. When the density is low at x the local bandwidths are increased to overcome the sparseness of data. On the other hand, when many data points are available around x the bandwidths can be reduced. Particularly, one could choose the local bandwidths such that only one observation is in the local neighbourhood. This would lead to the widely used *nearest neighbour matching* estimator or *pair-matching* estimator (Rubin 1974), where the local neighbourhood is often defined via the Mahalanobis distance.

The *exact matching* estimator, on the other hand, would be obtained by defining the kernel function as $\mathbf{K}_H(u) = 1$ for $u = 0$ and zero otherwise. As mentioned before, exact matching will often be very difficult to implement. If X contains continuous regressors, it will be literally speaking impossible to find observations with exactly the value x. But even if X contains only, say, 10 or 20 dummy variables, the number of cells spanned by these dummy variables is very large. For many values of x there may be no observation X_j that is equal to x in all respects, such that for those x the estimate would be undefined. Discarding these values x from the estimation (6), however, may lead to a very selected subpopulation such that results may not be representative for the population of interest. Even if for all values of x at least one alike observation in the $D = 0$ and the $D = 1$ subsample can be found, the procedure is likely to lead to a rather high variance.

The kernel matching estimator with *fixed* bandwidth gives non-zero weights to all observations within the local neighbourhood. When the dimension of X is relatively large it will often be more stable computationally if a kernel with infinite support is used such that $\mathbf{K}_H(X - x) \neq 0$, whatever the distance between X and x. In other words, all observations are used in the local averaging but with weights decreasing with distance to x. (On the other hand, this increases computation time particularly for local linear or local polynomial estimators considered below.)

As shown in Abadie and Imbens (2006a), nearest neighbour matching estimators are usually inconsistent unless the dimension of X is small (e.g., when the propensity score is used instead) or the number of nearest neighbours included increases asymptotically to infinity. In addition, the bootstrap may not work for nearest neighbour matching to obtain standard errors, see Abadie and Imbens (2006b). In contrast, kernel matching can achieve $\sqrt{n}$ consistent estimation and the bootstrap is suspected to work (Heckman et al. 1998b, Heckman et al. 1998a). Note that for achieving $\sqrt{n}$ consistency, higher order kernels are often required, implying that some observations receive negative kernel weights. These seem to be rarely used in applications, though.

Instead of kernel regression, *local linear regression* estimates $\hat{m}_0(x)$ as

$$\hat{m}_0(x) = \hat{\alpha} \qquad \text{where } (\hat{\alpha}, \hat{\beta}) = \arg\min_{\alpha,\beta} \sum_{j:D_j=0} \big(Y_j - \alpha - \beta(X_j - x)\big)^2 \cdot \mathbf{K}_H\big(X_j - x\big). \tag{7}$$

Local linear regression is known to have better properties in boundary regions (Fan 1992, Fan and Gijbels 1996), which can be relevant here if the densities of $f_{X|D=1}$ and $f_{X|D=0}$ are quite distinct.[6] If the outcome variable Y is binary or bounded, e.g., between zero and one, *local logit* regression can lead to a lower bias than local linear regression (Gozalo and Linton 2000). For details on local logit regression see e.g., Frölich (2006).

Regarding the kernel function $\mathbf{K}_H(\cdot)$, one should take into account that in many applications, the set of control variables X often contains continuous, discrete and binary variables. The *asymptotic* properties of matching estimators usually depend only on the continuous variables, and it is often suggested to conduct separate regressions for each cell defined by the discrete regressors. With a large number of X regressors and limited sample size this will often not be possible, and even if it were, it would often lead to imprecise estimates. Even with only 10 dummy variables, 2^{10} different cells would have to be accommodated. In many economic applications it appears to make sense to consider two observations which agree in, say, 9 out of 10 binary characteristics to be more similar than two observations which are equal on fewer characteristics. Indeed, smoothing over the dummy variables can improve substantially the precision in small samples, see Racine and Li (2004). Therefore, smoothing over the discrete and binary regressors will often be very important. To implement the multivariate kernel function $\mathbf{K}_H(X_j - x)$, a *product kernel* is often convenient and fast to implement, which also permits one to coalesce continuous and discrete regressors. We may distinguish three types of regressors: continuous, discrete with natural ordering (e.g., number of children) and discrete without natural ordering (e.g., blue, red, green). Suppose that the variables in X are arranged such that the first q_1 regressors are continuous, the regressors $q_1 + 1, \ldots, q_2$ discrete with natural ordering and the remaining $Q - q_2$ regressors discrete without natural ordering, including binary variables. Then the kernel weights $\mathbf{K}_H(X_j - x)$ are computed as

$$\mathbf{K}_{h,\delta,\lambda}(X_j - x) = \prod_{q=1}^{q_1} \kappa\left(\frac{X_{q,j} - x_q}{h}\right) \cdot \prod_{q=q_1+1}^{q_2} \delta^{|X_{q,j} - x_q|} \cdot \prod_{q=q_2+1}^{Q} \lambda^{1(X_{q,j} \neq x_q)}, \qquad (8)$$

where $X_{q,j}$ and x_q denote the q-th element of X_j and x, respectively, $1(\cdot)$ denotes the indicator function, κ is a symmetric *univariate* kernel function, e.g., the Gaussian kernel, and h, δ, and λ are positive bandwidth parameters with $0 \leq \delta, \lambda \leq 1$. This kernel function $\mathbf{K}_{h,\delta,\lambda}(X_j - x)$ measures the distance between X_j and x through three components: The first term is the conventional product kernel for continuous regressors with h defining the size of the local neighbourhood. The second term measures the distance between the ordered discrete regressors and assigns geometrically declining weights to more unlike observations. The third term measures the mismatch between the unordered discrete regressors. δ controls the amount of smoothing for the ordered and λ for the unordered discrete regressors. For example, the multiplicative weight contribution of the Q-th regressor is 1 if the Q-th element of X_j and of x are identical and λ if they are different. The larger δ and/or λ the more smoothing takes place,

[6] Local *higher order* polynomials may be computationally demanding if the dimension of X is high and could require rather large bandwidth values to avoid local collinearity. Local linear regression seems to be preferred in practice.

with respect to the discrete regressors. If δ and λ are both 1, the discrete regressors would not affect the kernel weights and the nonparametric estimator would 'smooth globally' over the discrete regressors. These variables would nevertheless still enter in the local hyperplane, i.e., in the second last term in (7). On the other hand, if δ and λ are both zero, smoothing would proceed only within each of the cells defined by the discrete regressors but not between them. If, further, X contained no continuous regressors this would correspond to the frequency estimator, where Y is estimated by the average of the observations within each cell.

Any values between 0 and 1 for δ and λ thus correspond to some smoothing over the discrete regressors. By noting that

$$\prod \lambda^{1(X_{q,j} \neq x_q)} = \lambda^{\sum 1(X_{q,j} \neq x_q)},$$

the weight contribution of the unordered discrete regressors thus depends only on the *number* of regressors that are distinct between X_j and x.

Principally, instead of using only three bandwidth values h, δ, λ for all regressors, a different bandwidth could be employed for each regressor. But this would increase substantially the computational burden for bandwidth selection and might lead to additional noise due to estimating these bandwidth parameters. Nevertheless, groups of similar regressors could be formed, with each group assigned a separate bandwidth parameter, if the explanatory variables are deemed too distinct. Particularly if the ranges assumed by the ordered discrete variables vary considerably, those variables that take on many different values should be separated from those with only few values.[7] Moreover, the continuous regressors should be rotated to have mean zero, variance one and zero correlation between each other.

To choose the bandwidth values, leave-one-out least-squares cross-validation is often applied, separately for the $D = 1$ and the $D = 0$ group. Cross-validation is known to be inconsistent for estimating ATE and ATET since some asymptotic undersmoothing would be required to achieve $\sqrt{n}$ consistency. The bandwidths obtained by cross-validation are usually too large such that bias is too large relative to the variance. Therefore, one would like to undersmooth with respect to the bandwidths obtained by cross-validation. On the other hand, for propensity score matching it turned out that cross-validation actually performed very well in a number of simulations, see e.g., Frölich (2004) or Frölich (2005). It is unclear, however, whether this finding would also hold for higher dimensional X, where more undersmoothing would be required.

For two recent applications of this methodology see e.g., Frölich and Michaelowa (2005) and Bourdon et al. (2007).

4 Conclusions

The previous derivations have shown that the semiparametric variance bound is lower for matching on X than for propensity score matching. Hence, matching estima-

[7] Alternatively, instead of geometrically declining weights, one could simply use the weights defined by the kernel function κ. Then the ordered discrete variables would be treated as the continuous variables.

tion on X should be considered as a serious alternative, although it is much more computationally demanding than PSM with a parametrically estimated propensity score.

The results obtained referred to a situation where the propensity score is known. If the propensity score is unknown, its estimation would often add even further to the variance of the PSM estimator, see Heckman et al. (1998b). In principle, it cannot be ruled out that the additional terms appearing in the asymptotic expansion of the PSM estimator due to the estimation of the propensity score $p(x)$ are strongly negatively correlated with the terms due to nonparametric estimation of $\mathfrak{m}_d(p(x))$. This, however, would require a very particular joint estimator of $p(x)$ and $\mathfrak{m}_d(p(x))$, where the nonparametric estimator $\mathfrak{m}_d(p(x))$ is modified to take the local bias and local variance of the propensity score estimator into account. Nevertheless, PSM with estimated propensity score can be more precise than PSM with the true propensity score in certain situations. This is e.g., the case when all the covariates X are discrete. (See also Rosenbaum and Rubin (1985), Rubin and Thomas (1992, 1996, 2000).) Hence, in those situations where PSM with estimated propensity score is more precise than PSM with the true propensity score, the main result of Sect. 2 may not apply. Otherwise, however, PSM is generally less precise in estimating ATE than matching on X. For ATET, the result is less clear, though, since not knowing the propensity score increases the variance bound $\mathcal{V}^X_{\text{ATET}}$ and may reverse the ranking, as found e.g., in Heckman et al. (1998b).[8]

A Proof of Theorem 1

With some straightforward calculations, given later, it can shown that

$$\acute{\sigma}_d(\rho) = E\left[\sigma_d^2(X)|p(X)=\rho\right] + Var\left[m_d(X)|p(X)=\rho\right] \tag{9}$$

and that for any constant a:

$$\begin{aligned} E\left[(m_1(X)-m_0(X)-a)^2 \;|p(X)=\rho\right] &= Var\left[m_1(X)-m_0(X)\,|p(X)=\rho\right] \\ &\quad + (\mathfrak{m}_1(\rho)-\mathfrak{m}_0(\rho)-a)^2\;. \end{aligned} \tag{10}$$

With these two preliminaries the difference $\mathcal{V}^{\text{PSM}}_{\text{ATE}} - \mathcal{V}^X_{\text{ATE}}$ can be written as

$$\begin{aligned} &\mathcal{V}^{\text{PSM}}_{\text{ATE}} - \mathcal{V}^X_{\text{ATE}} \\ &\quad = E\left[\frac{\acute{\sigma}_1^2(P)}{P} + \frac{\acute{\sigma}_0^2(P)}{1-P}\right] - E\left[\frac{\sigma_1^2(X)}{p(X)} + \frac{\sigma_0^2(X)}{1-p(X)}\right] \\ &\qquad + E\left[(\mathfrak{m}_1(P)-\mathfrak{m}_0(P)-\text{ATE})^2\right] \\ &\qquad - E\left[\,E\left[(m_1(X)-m_0(X)-\text{ATE})^2\;|p(X)=P\right]\right] \end{aligned}$$

[8] This is in contrast to propensity score weighting, where only $p(x)$ needs to be estimated and an efficient estimated propensity score weighting estimator has been proposed by Hirano et al. (2003). A deeper analysis is beyond the scope of this note.

$$
\begin{aligned}
&= E\left[\frac{Var\left[m_1(X)|p(X)=P\right]}{P}+\frac{Var\left[m_0(X)|p(X)=P\right]}{1-P}\right] \\
&\quad + E\left[(\mathfrak{m}_1(P)-\mathfrak{m}_0(P)-\mathrm{ATE})^2\right] \\
&\quad - E\left[\,Var\left[m_1(X)-m_0(X)\,|p(X)=P\right]+(\mathfrak{m}_1(P)-\mathfrak{m}_0(P)-\mathrm{ATE})^2\right] \\
&= E\left[\frac{V_1(P)}{P}+\frac{V_0(P)}{1-P}\right]-E\left[\,Var\left[m_1(X)-m_0(X)\,|p(X)=P\right]\right] \\
&= E\left[\frac{V_1(P)}{P}+\frac{V_0(P)}{1-P}\right]-E\left[V_1(P)+V_0(P)-2C(P)\right] \\
&= E\left[\frac{1-P}{P}V_1(P)+\frac{P}{1-P}V_0(P)+2C(P)\right],
\end{aligned}
$$

where $V_d(\rho)\equiv Var\left(m_d(X)|p(X)=\rho\right)$ and $C(\rho)\equiv cov\left(m_1(X),m_0(X)|p(X)=\rho\right)$.

The first two terms are positive but the covariance term could be negative. Since the covariance is bounded in absolute value by the variances: $|C(\rho)|\leq\sqrt{V_1(\rho)V_0(\rho)}$, a lower bound on $\mathcal{V}^{\mathrm{PSM}}_{\mathrm{ATE}}-\mathcal{V}^{X}_{\mathrm{ATE}}$ is

$$
\begin{aligned}
\mathcal{V}^{\mathrm{PSM}}_{\mathrm{ATE}}-\mathcal{V}^{X}_{\mathrm{ATE}} &\geq E\left[\frac{1-P}{P}V_1(P)+\frac{P}{1-P}V_0(P)-2\sqrt{V_1(P)V_0(P)}\right] \\
&= E\left[V_0(P)\left(\frac{1-P}{P}\frac{V_1(P)}{V_0(P)}+\frac{P}{1-P}-2\sqrt{\frac{V_1(P)}{V_0(P)}}\right)\right] \\
&= E\left[V_0(P)\frac{\left(\sqrt{\frac{V_1(P)}{V_0(P)}}(1-P)-P\right)^2}{P\,(1-P)}\right]\geq 0\,,
\end{aligned}
$$

which is non-negative. Generally, $\mathcal{V}^{\mathrm{PSM}}_{\mathrm{ATE}}-\mathcal{V}^{X}_{\mathrm{ATE}}$ is strictly positive unless the support of P contains *only* values where either both variances $V_1(\rho)$ and $V_0(\rho)$ are zero or where $\sqrt{\frac{V_1(\rho)}{V_0(\rho)}}=\frac{\rho}{1-\rho}$ *and* $corr\left(m_1(X),m_0(X)|p(X)=\rho\right)=-1$.

Analogously, for ATET it follows for $\mathcal{V}^{\mathrm{PSM}}_{\mathrm{ATET}}-\mathcal{V}^{X}_{\mathrm{ATET}}$

$$
\begin{aligned}
&\left(\mathcal{V}^{\mathrm{PSM}}_{\mathrm{ATET}}-\mathcal{V}^{X}_{\mathrm{ATET}}\right)\cdot\Gamma^2 \\
&= E\left[P\acute{\sigma}_1^2(P)+P^2\frac{\acute{\sigma}_0^2(P)}{1-P}\right]-E\left[p(X)\sigma_1^2(X)+p^2(X)\frac{\sigma_0^2(X)}{1-p(X)}\right] \\
&\quad + E\left[P^2\left(\mathfrak{m}_1(P)-\mathfrak{m}_0(P)-\mathrm{ATET}\right)^2\right] \\
&\quad - E\left[P^2\cdot E\left[(m_1(X)-m_0(X)-\mathrm{ATET})^2\;|p(X)=P\right]\right] \\
&= E\left[P\cdot V_1(P)+\frac{P^2}{1-P}V_0(P)\right]+E\left[P^2\left(\mathfrak{m}_1(P)-\mathfrak{m}_0(P)-\mathrm{ATET}\right)^2\right] \\
&\quad - E\left[P^2\cdot\left(Var\left[m_1(X)-m_0(X)\,|p(X)=P\right]+(\mathfrak{m}_1(P)-\mathfrak{m}_0(P)-\mathrm{ATET})^2\right)\right]
\end{aligned}
$$

$$
\begin{aligned}
&= E\left[P \cdot V_1(P) + \frac{P^2}{1-P} V_0(P)\right] - E\left[P^2 \cdot (Var\left[m_1(X) - m_0(X) \,|\, p(X) = P\right])\right] \\
&= E\left[P \cdot V_1(P) + \frac{P^2}{1-P} V_0(P)\right] - E\left[P^2 \cdot (V_1(P) + V_0(P) - 2C(P))\right] \\
&= E\left[P^2 \cdot \left(\frac{1-P}{P} V_1(P) + \frac{P}{1-P} V_0(P) + 2C(P)\right)\right] \\
&\geq E\left[P^2 \cdot \left(\frac{1-P}{P} V_1(P) + \frac{P}{1-P} V_0(P) - 2\sqrt{V_1(P) V_0(P)}\right)\right] \\
&= E\left[P^2 V_0(P) \left(\frac{1-P}{P} \frac{V_1(P)}{V_0(P)} + \frac{P}{1-P} - 2\sqrt{\frac{V_1(P)}{V_0(P)}}\right)\right] \\
&= E\left[P^2 \cdot V_0(P) \frac{\left(\sqrt{\frac{V_1(P)}{V_0(P)}}(1-P) - P\right)^2}{P(1-P)}\right] \geq 0,
\end{aligned}
$$

and the same conclusions apply.

It remains to prove (9) and (10). Consider (9) first. The following calculations show that

$$
\acute{\sigma}_d(\rho) = E\left[\sigma_d^2(X) | p(X) = \rho\right] + Var\left[m_d(X) | p(X) = \rho\right],
$$

because

$$
\begin{aligned}
\acute{\sigma}_d(\rho) &= Var[Y | p(X) = \rho, D = d] = E\left[(Y - \mathfrak{m}_d(\rho))^2 \,|\, p(X) = \rho, D = d\right] \\
&= E\left[E\left[(Y - \mathfrak{m}_d(\rho))^2 \,|\, X, D = d\right] | p(X) = \rho, D = d\right] \\
&= E\left[E\left[(Y - m_d(X) + m_d(X) - \mathfrak{m}_d(\rho))^2 \,|\, X, D = d\right] | p(X) = \rho, D = d\right] \\
&= E\left[E\left[(Y - m_d(X))^2 \,|\, X, D = d\right] | p(X) = \rho, D = d\right] \\
&\quad + E\left[E\left[(m_d(X) - \mathfrak{m}_d(\rho))^2 \,|\, X, D = d\right] | p(X) = \rho, D = d\right] \\
&\quad + E\left[E\left[(Y - m_d(X))(m_d(X) - \mathfrak{m}_d(\rho)) \,|\, X, D = d\right] | p(X) = \rho, D = d\right] \\
&= E\left[\sigma_d^2(X) + (m_d(X) - \mathfrak{m}_d(\rho))^2 \,|\, p(X) = \rho, D = d\right] \\
&= \int \left\{\sigma_d^2(X) + (m_d(X) - \mathfrak{m}_d(\rho))^2\right\} \cdot dF(X | p(X) = \rho, D = d) \\
&= E\left[\sigma_d^2(X) + (m_d(X) - \mathfrak{m}_d(\rho))^2 \,|\, p(X) = \rho\right],
\end{aligned}
$$

where the last equality holds because the distribution of X is independent of D given $p(X)$. To see this notice that D is binary and therefore $X \perp\!\!\!\perp D | p(X)$ is equivalent to saying $\Pr(D = 1 | X, p(X)) = \Pr(D = 1 | p(X))$, which is true by the definition of the propensity score.

Now consider (10). It is shown that for any constant a:

$$E\left[(m_1(X)-m_0(X)-a)^2 \mid p(X)=\rho\right] = Var\left[m_1(X)-m_0(X) \mid p(X)=\rho\right] + (\mathfrak{m}_1(\rho)-\mathfrak{m}_0(\rho)-a)^2,$$

because

$$\begin{aligned}
&E\left[(m_1(X)-m_0(X)-a)^2 \mid p(X)=\rho\right] \\
&= E\left[((m_1(X)-m_0(X)-\mathfrak{m}_1(\rho)-\mathfrak{m}_0(\rho))+(\mathfrak{m}_1(\rho)-\mathfrak{m}_0(\rho)-a))^2 \mid p(X)=\rho\right] \\
&= E\left[(m_1(X)-m_0(X)-\mathfrak{m}_1(\rho)-\mathfrak{m}_0(\rho))^2 \mid p(X)=\rho\right] \\
&\quad + E\left[(\mathfrak{m}_1(\rho)-\mathfrak{m}_0(\rho)-a)^2 \mid p(X)=\rho\right] \\
&\quad + 2E\left[(m_1(X)-m_0(X)-\mathfrak{m}_1(\rho)-\mathfrak{m}_0(\rho))\,(\mathfrak{m}_1(\rho)-\mathfrak{m}_0(\rho)-a) \mid p(X)=\rho\right] \\
&= Var\left[m_1(X)-m_0(X) \mid p(X)=\rho\right] + (\mathfrak{m}_1(\rho)-\mathfrak{m}_0(\rho)-a)^2 .
\end{aligned}$$

Acknowledgement I would like to thank the editors and two anonymous referees for helpful comments.

References

Abadie, A., Imbens, G. (2006a) Large sample properties of matching estimators for average treatment Effects. Econometrica **74**, 235–267

Abadie, A., Imbens, G. (2006b) On the failure of the bootstrap for matching estimators. NBER Technical Working Paper No. 325

Begun, J., Hall, W., Huang, W., Wellner, J. (1983) Information and asymptotic efficiency in parametric-nonparametric models. Annals of Statistics **11**, 432–452

Bickel, P., Klaassen, C., Ritov, Y., Wellner, J. (1993) Efficient and Adaptive Estimation for Semiparametric Models. John Hopkins University Press, Baltimore

Black, D., Smith, J. (2004) How robust is the evidence on the effects of college quality? Evidence from matching. Journal of Econometrics **121**, 99–124

Bourdon, J., Frölich, M., Michaelowa, K. (2007) Teacher Shortages, Teacher Contracts and their Impact on Education in Africa. IZA Discussion paper 2844

Fan, J. (1992) Design-adaptive nonparametric regression. Journal of American Statistical Association **87**, 998–1004

Fan, J., Gijbels, I. (1996) Local Polynomial Modeling and its Applications. Chapman and Hall, London

Frölich, M. (2004) Finite sample properties of propensity-score matching and weighting estimators. The Review of Economics and Statistics **86**, 77–90

Frölich, M. (2005) Matching estimators and optimal bandwidth choice. Statistics and Computing **15**(3), 197–215

Frölich, M. (2006) Nonparametric regression for binary dependent variables. Econometrics Journal **9**, 511–540

Frölich, M. (2007) Propensity score matching without conditional independence assumption – with an application to the gender wage gap in the UK. Econometrics Journal **10**, 359–407

Frölich, M., Heshmati, A., Lechner, M. (2004) A microeconometric evaluation of rehabilitation of long-term sickness in Sweden. Journal of Applied Econometrics **19**, 375–396

Frölich, M., Michaelowa, K. (2005) Peer effects and textbooks in primary education: evidence from Francophone Sub-Saharan Africa. IZA Discussion paper 1519

Gerfin, M., Lechner, M. (2002) Microeconometric evaluation of the active labour market policy in Switzerland. Economic Journal **112**, 854–893

Gozalo, P., Linton, O. (2000) Local nonlinear least squares: Using parametric information in nonparametric regression. Journal of Econometrics **99**, 63–106

Hahn, J. (1998) On the role of the propensity score in efficient semiparametric estimation of average treatment effects. Econometrica **66**, 315–331
Heckman, J., Ichimura, H., Smith, J., Todd, P. (1998a) Characterizing selection bias using experimental data. Econometrica **66**, 1017–1098
Heckman, J., Ichimura, H., Todd, P. (1998b) Matching as an econometric evaluation estimator. Review of Economic Studies **65**, 261–294
Hirano, K., Imbens, G., Ridder, G. (2003) Efficient estimation of average treatment effects using the estimated propensity score. Econometrica **71**, 1161–1189
Koshevnik, Y., Levit, B. (1976) On a non-parametric analogue of the information matrix. Theory of Probability and Applications **21**, 738–753
Larsson, L. (2003) Evaluation of swedish youth labour market programmes. Journal of Human Resources **38**, 891–927
Lechner, M. (1999) Earnings and employment effects of continuous off-the-job training in East Germany after unification. Journal of Business and Economic Statistics **17**, 74–90
Lechner, M. (2002) Program heterogeneity and propensity score matching: An application to the evaluation of active labor market policies. The Review of Economics and Statistics **84**, 205–220
Newey, W. (1990) Semiparametric efficiency bounds. Journal of Applied Econometrics **5**, 99–135
Newey, W. (1994) The asymptotic variance of semiparametric estimators. Econometrica **62**, 1349–1382
Pfanzagl, J., Wefelmeyer, W. (1982) Contributions to a general asymptotic statistical theory. Springer, Heidelberg
Racine, J., Li, Q. (2004) Nonparametric estimation of regression functions with both categorical and continuous data. Journal of Econometrics **119**, 99–130
Rosenbaum, P., Rubin, D. (1983) The central role of the propensity score in observational studies for causal effects. Biometrika **70**, 41–55
Rosenbaum, P., Rubin, D. (1985) The bias due to incomplete matching. Biometrics **41**, 103–116
Rubin, D. (1974) Estimating causal effects of treatments in randomized and nonrandomized studies. Journal of Educational Psychology **66**, 688–701
Rubin, D., Thomas, N. (1992) Affinely invariant matching methods with ellip-soidal distributions. Annals of Statistics **20**, 1079–1093
Rubin, D., Thomas, N. (1996) Matching using estimated propensity scores: relating theory to practice. Biometrics **52**, 249–264
Rubin, D., Thomas, N. (2000) Combining propensity score matching with additional adjustments for prognostic covariates. Journal of the American Statistical Association **95**, 573–585
Sianesi, B. (2004) An evaluation of the swedish system of active labor market programs in the 1990s. The Review of Economics and Statistics **86**, 133–155
Smith, J., Todd, P. (2005) Does matching overcome LaLonde's critique of nonexperimental estimators? Journal of Econometrics **125**, 305–353
Stein, C. (1956) Efficient nonparametric testing and estimation. In: Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, vol. 1. University of California Press, Berkeley